Submitted (2019), **0**, 0, *pp.* **??**–30
doi:xxxxx/submitted/maxcombo-cox-bio1

On weighted log-rank combination tests and companion Cox model estimators

LARRY F. LEÓN*, RAY LIN

Department of Biostatistics, Genentech, California, U.S.A.

KEAVEN M. ANDERSON

Late Development Statistics, Merck Research Laboratories, Pennsylvania, U.S.A.

larry.leon.05@post.harvard.edu

Summary

In randomized clinical trials the log-rank test and Cox proportional hazards model are the gold standard in survival data analyses. While the log-rank test is generally valid, in the presence of non-proportional hazards the power can be substantially decreased relative to the proportional hazards assumptions under which studies are usually designed. In contrast, weighted log-rank tests can be more powerful for specific treatment differences under non-proportional hazards scenarios; However, a poor choice of the weighting form can be detrimental. Recent work on combining various weighted log-rank tests allows for tests that are capable of detecting treatment effects across a broad range of non-proportional hazards scenarios. In addition, we describe how tests based on restricted mean survival time comparisons can be included within combinations of weighted log-rank tests (as well as other test statistics such as Tarone-Ware and Renyi-type supremum families). For estimation, we propose companion weighted Cox model estimators (Lin

*To whom correspondence should be addressed.

(1991), and Sasieni (1993)). The companion Cox estimators utilize the weighting form that is "selected" through the combination test. A versatile resampling procedure (Dobler *and others*, 2017) is employed which allows for the aforementioned combinations across various test statistics and as a by-product, simultaneous confidence intervals for the companion estimators (e.g., across weighted Cox models and RMST). In clinical settings such as the recent immunotherapy oncology trials which exhibit consistent patterns of late-separation the proposed methods may serve as complementary summaries to the standard log-rank and Cox model analyses. In particular, in late-separation settings where there is, on average, lack of early benefit followed by substantial later benefit, the standard Cox estimator is generally neither accurate for lack of early benefit nor accurate for the later term benefit; Whereas weighted Cox model estimators can provide more accurate summaries of such delayed effects. The performance of various combinations and their companion Cox estimators as well as RMST are evaluated in simulation studies under null, proportional hazards, late-separation, and early-separation scenarios. For illustration we apply the proposals to a randomized clinical trial study of the PD-L1-targeted therapy atezolizumab in comparison with docetaxel in previously treated non-small-cell lung cancer patients. R code can be found on https://github.com/larry-leon/combination-tests-and-estimators.

Key words: Weighted Cox model; Non-proportional hazards; Weighted-log rank test; Combination test; Restricted mean survival time (RMST).

1. Introduction

In randomized clinical trials the log-rank test and Cox model estimator are the gold standard primary analyses for survival data. While the log-rank test is generally valid, in the presence of non-proportional hazards the power can be substantially decreased relative to the proportional hazards assumptions under which studies are usually designed. In contrast, weighted log-rank tests can be more powerful for specific treatment differences under non-proportional hazards scenarios; For example, the class of $G^{\rho,\gamma}$ weights (Fleming and Harrington, 1991) allows for a wide range of weighting patterns that can focus on specific forms of departures such as late-separation and early-separation scenarios. In particular, the $G^{0,1}$ weighting scheme which we denote as FH(0,1) in the sequel, can be more powerful in detecting late-separation patterns.

As an illustration we consider the recent OAK clinical trial study (Rittmeyer *and others*, 2017) which was an open-label randomized trial evaluating the PD-L1-targeted therapy atezolizumab in comparison with docetaxel in previously treated non-small-cell lung cancer patients. The primary analysis was based on the first $n = 850$ randomized patients with a 1-sided alpha of 1.5% used for the overall population (while 1-sided 1% was used for the PD-L1 TC1/2/3 or IC1/2/3 subgroup). The study had more than 95% power under the study assumptions for the overall comparison at $\alpha = 1.5\%$. The primary analysis was specified to be conducted when approximately 70% of the patients had died. Between March 11, 2014, and Nov 28, 2014, 850 patients in the primary analysis population were recruited and the primary analysis was conducted based on the data cutoff on July 7, 2016.

Here we consider the hypothetical question of whether an earlier analysis could have concluded superiority. Figure 1 displays the Kaplan-Meier estimates for the treatment groups based on the July 7, 2016 cuttoff (denoted “final”) and interim data based on the retrospective cutoff of June 7, 2015 (denoted “interim” which is 13 months earlier). Suppose that for the total $\alpha = 1.5\%$ this was split at 0.5% for the interim and 1% for the final: That is, in order for the interim analysis to be considered significant the 1-sided p-value for the interim analysis would need to be < 0.005. Now, the log-rank p-value based on the interim data is 0.00501 which does not technically meet the 0.005 criteria; Whereas, the FH(0,1) p-value is 0.00179. Of course this is a hypothetical consideration, nevertheless, the potential advantage of weighted log-rank tests over standard log-rank analyses can be practically important when non-proportional hazards may be present.

In general it is not clear how to pre-specify an optimal weighting scheme and the potential power loss from a poorly chosen weighting scheme can be detrimental. To avoid the need to rely on a single choice, approaches based on combining various tests have recently been developed. In this paper we propose companion weighted Cox model estimators as well as describe how tests based on restricted mean survival time (RMST) comparisons can be included within combinations of weighted log-rank tests.

This paper is organized as follows. Section 2 reviews the combination of weighted log-rank tests as well as the option to include the RMST test. While the asymptotic distributions of combination tests are generally complex, we describe a relatively simple resampling approach that can be used across various combinations. Section 3 describes the proposed companion weighted Cox model as well as an example calculation for the asymptotic limiting value under non-proprotional hazards scenarios. Sections 4 and 5 provide simulations and additional details on the hypothetical interim analysis of the aforementioned OAK study. We conclude with a discussion in Section 6 and technical details are outlined in the Supplementary Material Section.

2. Maximum Combination Test

We consider the two-sample random censorship model and use the following notation for samples $j = 0, 1$. For $i = 1, \ldots, n$ let $Z_i = j$ if observation i is from group j and $n = n_0 + n_1$, with n_j the number of observations from group j. Let T denote the survival time and C the censoring time. We observed the possibly censored survival time $X = \min(T, C)$ and $\Delta = I(T \leqslant C)$ the event indicator. The survival times T are assumed to be independent of C conditional on Z. The triplets (Z_i, X_i, Δ_i) for $i = 1, \ldots, n$ are assumed to be iid replicates.

The at-risk and counting processes for samples $j = 0, 1$ are $Y_{i,j}(t) = I(X_i \geqslant t, Z_i = j)$ and $N_{i,j}(t) = I(X_i \leqslant t, \Delta_i = 1, Z_i = j)$. Define $\bar{Y}_j(t) = \sum_{i=1}^{n} Y_{i,j}(t)$, $\bar{Y}(t) = \bar{Y}_0(t) + \bar{Y}_1(t)$,

$\bar{N}_j(t) = \sum_{i=1}^{n} N_{i,j}(t)$, and $\bar{N}(t) = \bar{N}_0(t) + \bar{N}_1(t)$.

Let F_j denote the distribution functions for survival times from group $j = 0, 1$ and $\bar{F}_j = 1 - F_j$. Define the pooled versions analogously (F and $\bar{F} = 1 - F$, say). The Kaplan-Meier estimators for the group survival distributions are denoted by $\hat{\bar{F}}_j$ (groups $j = 0, 1$) and for the pooled population by $\hat{\bar{F}}$.

The weighted log-rank test of class K^+ described by Gill (1980) is

$$W_K = \int_0^\infty \left\{ \frac{K(t)}{\bar{Y}_0(t)} d\bar{N}_0(t) - \frac{K(t)}{\bar{Y}_1(t)} d\bar{N}_1(t) \right\} \tag{2.1}$$

where $K(t) = \sqrt{\frac{n_0+n_1}{n_0 n_1}} w(t) \frac{\bar{Y}_0(t)\bar{Y}_1(t)}{\bar{Y}(t)}$ with $w(t)$ a non-negative bounded predictable process. In the sequel we will focus on the family of $G^{\rho,\gamma}$ weighted log-rank tests denoted hereafter as FH(ρ, γ) with weights $w(t) = [\hat{\bar{F}}(t-)]^\rho [1 - \hat{\bar{F}}(t-)]^\gamma$ where $\rho \geqslant 0$, $\gamma \geqslant 0$, and $\hat{\bar{F}}(t-)$ is the left-continuous version of the Kaplan-Meier estimator for the pooled sample.

The FH(ρ, γ) weights allow for the construction of tests that can have enhanced ability to detect specific treatment differences. For example, Figure 2 displays underlying Kaplan-Meier population curves for the late-separation setting described in Section 3 as well as the weight profiles for FH(0,1), FH(1,0), and FH(1,1) weighting. These weighting schemes "focus" on late-separation, early-separation, and middle-separation patterns, respectively. In this late-separation scenario, the weighted log-rank test based on FH(0,1) will generally be more powerful than log-rank, FH(1,0), and FH(1,1) tests.

Now, the asymptotic null distribution of W_K is normal with mean zero and variance σ_K^2 (say) which can be consistently estimated by (Fleming and Harrington, 1991)

$$\hat{\sigma}_K^2 = \int_0^\infty \left\{ \frac{K(t)^2}{\bar{Y}_0(t)} + \frac{K(t)^2}{\bar{Y}_1(t)} \right\} \left\{ 1 - \frac{\Delta\bar{N}(t) - 1}{\bar{Y}(t) - 1} \right\} \frac{d\bar{N}(t)}{\bar{Y}(t)}, \tag{2.2}$$

where $\Delta\bar{N}(t) = \bar{N}(t) - \bar{N}(t-)$. The 1-sided Z-statistic (superiority test for experimental) is then $Z_K = W_K / \hat{\sigma}_K$.

In the sequel, let $Z_1, Z_2, \ldots, Z_q$ denote q tests based on (q different) FH(ρ, γ) weighting schemes and we are interested in the combination test $Z_{max} := \max(Z_1, Z_2, ..., Z_q)$ which takes the maximum of the q z-statistics. For example, if combining two weighting schemes FH(ρ_1, γ_1) and FH(ρ_2, γ_2), then Z_1 denotes $Z_{K_1} = W_{K_1}/\hat{\sigma}_{K_1}$ with $K_1(t) = \sqrt{\frac{n_0+n_1}{n_0 n_1}}[\hat{\bar{F}}(t-)]^{\rho_1}[1-\hat{\bar{F}}(t-)]^{\gamma_1} \frac{\bar{Y}_0(t)\bar{Y}_1(t)}{\bar{Y}(t)}$ and Z_2 is defined analogously with $K_2(t) = \sqrt{\frac{n_0+n_1}{n_0 n_1}}[\hat{\bar{F}}(t-)]^{\rho_2}[1 - \hat{\bar{F}}(t-)]^{\gamma_2} \frac{\bar{Y}_0(t)\bar{Y}_1(t)}{\bar{Y}(t)}$.

In addition to combining weighted log-rank tests we also consider including the restricted mean survival time (RMST) test (Royston and Parmar, 2016; Zhao *and others*, 2016) within the combination. We note that Chi and Tsai (2001) considered combinations of weighted log-rank and weighted Kaplan-Meier tests (Pepe and Fleming, 1989) and provide large sample approximations to calculating p-values based on estimated correlations between tests.

Here we consider the RMST test based on the truncation point $\tau_m = \max(\tau_0, \tau_1)$ where the τ_j are the largest non-censored (event) survival times for the control ($j = 0$) and treatment ($j = 1$) groups. Let $\hat{m}_j := \int_0^{\tau_m} \hat{\bar{F}}_j(t)dt$ denoted the RMST for group j. The variance of $\hat{m}_j$ can be estimated (see Corollary 3.2 of Gill (1983)) as follows: Define $\tilde{m}_j(t) = \hat{m}_j - \hat{m}_j(t)$ where $\hat{m}_j(t) := \int_0^t \hat{\bar{F}}_j(s)ds$. Then the variance of $\hat{m}_j$, σ_j^2 (say), can be consistently estimated by $\hat{\sigma}_j^2 = \int_0^{\tau_m} \tilde{m}_j(t)^2 \frac{d\bar{N}_j(t)}{\bar{Y}_j(t)(\bar{Y}_j(t)-1)}$. The z-statistic based on the RMST difference based on truncation point τ_m is defined as $Z_{d(\tau_m)} = \hat{d}/\hat{\sigma}_d$ where $\hat{d} = \hat{m}_1 - \hat{m}_0$ and $\hat{\sigma}_d^2 = \hat{\sigma}_1^2 + \hat{\sigma}_0^2$.

In the following we describe a resampling approach that can be used for tests based on combining across various types of test statistics such as weighted log-rank, RMST, and Renyi-type supremum versions of weighted log-rank tests (Fleming *and others*, 1987; Eng and Kosorok, 2005).

For each weighted log-rank test, under the null H_0, we have

$$W_K = \int_0^\infty \frac{K(t)}{\bar{Y}_0(t)} d\bar{M}_0(t) - \int_0^\infty \frac{K(t)}{\bar{Y}_1(t)} d\bar{M}_1(t),$$

where $\bar{M}_j(t) = \sum_{i=1}^n M_{i,j}(t)$ are martingale processes for the control ($j = 0$) and experimental groups ($j = 1$); $M_{i,j}(t) = N_{i,j}(t) - \int_0^t Y_{i,j}(s)\lambda_j(s)ds$ with λ_j the hazard function for group

j (Fleming and Harrington, 1991). Now, while there are available asymptotic approaches for calculating the p-values for tests based on combining a family of FH(ρ, γ) test statistics (Fleming and Harrington, 1991; Chi and Tsai, 2001; Lee, 2007) as noted by Dobler *and others* (2017) there can be advantages to resampling approaches in small to moderate samples (as with the bootstrap). We employ a resampling approach that has been used in several settings to approximate the distributions of complex survival analysis procedures (Dobler *and others*, 2017; Lin *and others*, 1993; Lin, 1997; Kosorok and Lin, 1999; Goldwasser *and others*, 2004; Tian *and others*, 2004). Moreover, while asymptotics are available for combining weighted log-rank tests, in this paper we also consider combinations of weighted log-rank and RMST tests.

The idea is to replace the unobservable martingale increments $dM_{i,j}$ in the above representation with independent standard normal realizations (independent of the data) that are multplied by the observable counting process increments $dN_{i,j}$ with unknown parameters "plugged-in" by consistent estimators (Dobler *and others*, 2017).

We first describe how this approach can be used to calculate the p-value for the combination test $Z_{max} = \max(Z_1, Z_2)$. Generate the processes $W_1^\dagger =$

$$\sum_{i=1}^{n} \int_0^\infty \left\{ \frac{K_1(t)}{\bar{Y}_0(t)} dN_{i,0}(t) \right\} G_{i,0} - \sum_{i=1}^{n} \int_0^\infty \left\{ \frac{K_1(t)}{\bar{Y}_1(t)} dN_{i,1}(t) \right\} G_{i,1}, \tag{2.3}$$

and $W_2^\dagger = \sum_{i=1}^n \int_0^\infty \left\{ \frac{K_2(t)}{\bar{Y}_0(t)} dN_{i,0}(t) \right\} G_{i,0} - \sum_{i=1}^n \int_0^\infty \left\{ \frac{K_2(t)}{\bar{Y}_1(t)} dN_{i,1}(t) \right\} G_{i,1}$,
where $G_{i,j}$ $(G_{1,0}, \ldots, G_{n_0,0}, G_{n_0+1,1}, \ldots, G_{1,n}$, say) are n i.i.d. $N(0,1)$ random variables that are independent of $Y_{i,j}$ and $N_{i,j}$. Then calculate $Z_1^\dagger = W_1^\dagger/\hat{\sigma}_1$, $Z_2^\dagger = W_2^\dagger/\hat{\sigma}_2$, and $Z_{max}^\dagger = \max(Z_1^\dagger, Z_2^\dagger)$. Repeat these steps a large number of times and then compare the observed value of $z_{max} = \max(z_1, z_2)$ to the empirical distribution of $Z_{max}^\dagger$.

Kosorok and Lin (1999) employ a similar approach for simulating their tests for which asymptotic approximations are not available; They term processes of the above forms "artificial martingales".

The extension to combining across a family of multiple FH(ρ, γ) members follows completely analagously. We note that the correlation is induced because the same $N(0,1)$ random variables are utilized for each group ($j = 0,1$) across the test statistics.

For the RMST test we have the following martingale representation for $\hat{\bar{F}}_j(t)$ (Fleming and Harrington, 1991):

$$(\hat{\bar{F}}_j(t) - \bar{F}_j(t)) \approx -\bar{F}_j(t)\int_0^t \frac{1}{\bar{Y}_j(s)} d\bar{M}_j(s), \text{ for } j = 0,1.$$

Now, using the same $G_{i,j}$ realizations as for $W_1^\dagger$ and $W_2^\dagger$ above, define the processes:

$$\nu_1^\dagger(t) = -\hat{\bar{F}}_1(t)\int_0^t \frac{1}{\bar{Y}_1(s)}\sum_{i=1}^n dN_{i,1}(s)G_{i,1};$$

$$\nu_0^\dagger(t) = -\hat{\bar{F}}_0(t)\int_0^t \frac{1}{\bar{Y}_0(s)}\sum_{i=1}^n dN_{i,0}(s)G_{i,0}.$$

The distribution of the centered RMST process can be approximated by $\{(\hat{m}_1 - \hat{m}_0) - (m_1 - m_0)\} :=$

$$\int_0^{\tau_m}\left\{(\hat{\bar{F}}_1(t) - \hat{\bar{F}}_0(t)) - (\bar{F}_1(t) - \bar{F}_0(t))\right\}dt \approx \int_0^{\tau_m}(\nu_1^\dagger(t) - \nu_0^\dagger(t))dt.$$

Under the null, $m_1 - m_0 = \int_0^{\tau_m}(\bar{F}_1(t) - \bar{F}_0(t))dt = 0$. Define $\hat{d}^\dagger = \int_0^{\tau_m}(\nu_1^\dagger(t) - \nu_0^\dagger(t))dt$ and form $Z_{d(\tau_m)}^\dagger = \hat{d}^\dagger/\hat{\sigma}_d$. Then the distribution of the test combining the two aforementioned weighted log-rank tests along with the RMST test is approximated by the empirical distribution of $Z_{max}^\dagger = \max(Z_1^\dagger, Z_2^\dagger, Z_{d(\tau_m)}^\dagger)$.

We note that Renyi-type supremum tests (Fleming *and others*, 1987; Eng and Kosorok, 2005) can also be included in the combination by proceeding in an analogous manner (As these are two-sided tests, any other tests included in the combination would also need to be two-sided.).

In the sequel we refer to this resampling approach as "synthetic-martingale" resampling.

3. Weighted Cox Model Estimators

We review the weighted Cox model estimation approach proposed by Lin (1991), and Sasieni (1993). Here we consider the FH(ρ,γ) weighted Cox model estimate that corresponds to the maximum among the FH(ρ,γ) weighted log-rank tests. That is, if among the q FH(ρ,γ) weighted log-rank tests, FH(ρ_l,γ_l) corresponds to the largest z-statistic, then we select the FH(ρ_l,γ_l)-weighted Cox model. We refer to this estimator as $\hat{\beta}_{max}$. Since the proposed companion weighted Cox model estimator $\hat{\beta}_{max}$ corresponds to a selection process we develop valid confidence intervals by constructing simultaneous confidence intervals that hold (simultaneously) across all weighting schemes included in the combination test.

As in the previous section we consider the two-sample problem where recall Z denotes the treatment group ($Z = 1$ for treated, and zero otherwise), $X = \min(T, C)$ denotes the possibly censored survival time and $\Delta = I(T \leqslant C)$ the event indicator. Here for the proportional hazards setup, let $Y_i(t) = I(X_i \geqslant t)$ and $N_i(t) = I(X_i \leqslant t, \Delta_i = 1)$ denote the at-risk and counting processes for subjects $i = 1, \ldots, n$.

Let $\lambda^h(t|Z)$ denote the true conditional hazard function and denote the corresponding true distribution by $F(t|Z)$ with density $f(t|Z)$. The censoring time C is assumed to be independent of T, conditional on Z, with conditional distribution $G(t|Z)$. Also, let $V(z)$ denote the distribution of the covariates with density $v(z)$ and define $\bar{F}(t|z) = 1 - F(t|z)$ and $\bar{G}(t|z) = 1-G(t|z)$. We use the following notation. Let $S^{(p)}(t) = n^{-1}\sum_{i=1}^n Z_i^p Y_i(t)\lambda^h(t|Z_i)$ and $S^{(p)}(t,\beta) = n^{-1}\sum_{i=1}^n Z_i^p Y_i(t)\exp(Z_i\beta)$, for $p = 0, 1$. In addition, let $s^{(p)}(t) = ES^{(p)}(t)$ and $s^{(p)}(t,\beta) = ES^{(p)}(t,\beta)$ denote their expectations taken under the true model.

The FH(ρ,γ)-weighted partial-likelihood is given by

$$l_w(\beta) = \sum_{i=1}^n \int_0^\infty w(t)\beta Z_i dN_i(t) - \int_0^\infty w(t)\log(S^{(0)}(t,\beta))\sum_{i=1}^n dN_i(t),$$

where $w(t)$ denotes the FH(ρ,γ) weights. Let $\hat{\beta}_w$ denote the maximizer of $l_w(\beta)$ which corresponds to the solution of

$$U_w(\beta) := \sum_{i=1}^n \int_0^\infty w(t)\left\{Z_i - \frac{S^{(1)}(t,\beta)}{S^{(0)}(t,\beta)}\right\} dN_i(t) = 0. \tag{3.4}$$

In the Supplementary Material Section we describe the details for estimating $\hat{\beta}_w$ as well as variance estimation. Moreover, because $\hat{\beta}_{max}$ based on FH(ρ_l, γ_l) weights corresponds to a selection process via the respective combination test, we also desribe in the Supplementary Material Section a resampling approach for calculating simultaneous confidence intervals for the collection of q FH(ρ, γ) weighted estimators. That is, let $\hat{\beta}_{w_1}/\hat{\sigma}_{w_1}, \ldots, \hat{\beta}_{w_q}/\hat{\sigma}_{w_q}$ denote the standardize weighted estimators. Then for c_{max} such that

$$\Pr\left\{\max\left\{|\hat{\beta}_{w_1}/\hat{\sigma}_{w_1}|, \ldots, |\hat{\beta}_{w_q}/\hat{\sigma}_{w_q}|\right\} \leqslant c_{max}\right\} \approx 1 - \alpha, \tag{3.5}$$

the confidence intervals $\hat{\beta}_{w_1} \pm \hat{\sigma}_{w_1} c_{max}$,...,$\hat{\beta}_{w_q} \pm \hat{\sigma}_{w_q} c_{max}$ will have approximate $(1-\alpha)\%$ coverage simultaneously, and in particular for $\hat{\beta}_{w_l} \pm \hat{\sigma}_{w_l} c_{max}$ corresponding to $\hat{\beta}_{max}$. A Wald-type test is therefore based on concluding superiority if the upper bound is < 0. In addition, if the RMST test is included then a simultaneous confidence interval across all estimators is constructed by including the standardized RMST estimator in terms of $|\hat{d}/\hat{\sigma}_d|$ in the above calculation of c_{max}.

In the following we briefly outline calculations for the asymptotic limit of the weighted estimator $\hat{\beta}_w$ under misspecification (Andersen and Gill, 1982; Struthers and Kalbfleisch, 1986; Lin, 1991) for a non-proportional hazards scenario in which a standard Cox model is misspecified in the presence of a late-separation effect. Under regularity conditions (Lin, 1991) $\hat{\beta}_w$ converges in probability to β^*_w which is the solution to $u_w(\beta) = 0$ (the limit, in probability, of $U_w(\beta)$) where

$$u_w(\beta) = \int_0^\infty w(t) \left\{ \frac{s^{(1)}(t)}{s^{(0)}(t)} - \frac{s^{(1)}(t, \beta)}{s^{(0)}(t, \beta)} \right\} s^{(0)}(t) dt. \tag{3.6}$$

Some preliminary calculations will be useful for deriving the form of $u_w(\beta)$ in specific settings. In particular, we will calculate β^*_w for the Cox estimator and various FH(ρ, γ) weighting schemes under a non-proportional hazards setting where the true model is $\lambda^h(t|z) = \lambda_0(t)\exp(\beta_0(t)z)$ with $\beta_0(t)$ denoting a time-varying effect; Whereas, the assumed misspecified model is a standard Cox model with treatment group indicator Z.

Now, the conditional probability of dying by time t given $Z = z$ without being censored is given by

$$\begin{aligned} Q_{X,\Delta}(t,1|z) &= \Pr(X \leqslant t, \Delta = 1|Z = z) \\ &= E_T I(s \leqslant t) \Pr(C \geqslant s|Z = z, T = s) \\ &= \int_0^t \bar{G}(s|z)\bar{F}(s|z)\lambda^h(s|z)ds, \end{aligned}$$

with density $q_{X,\Delta}(t,1|z) = \bar{G}(t|z)\bar{F}(t|z)\lambda^h(t|z)$. The joint density of (Z, X, Δ) with $\Delta = 1$ is then

$$q_{Z,X,\Delta}(z,t,1) = \bar{G}(t|z)\bar{F}(t|z)\lambda^h(t|z)v(z).$$

Note that $E(Y(t)|Z = z) = \Pr(X \geqslant t|Z = z) = \bar{G}(t|z)\bar{F}(t|z)$ and let $r(z,t,\beta) = \exp(z\beta)/\lambda^h(t|z)$ denote the ratio of the specified hazard link, $\exp(z\beta)$, to the true hazard $\lambda^h(t|z)$ (If the model is correctly specified, then evaluated at the true $\beta = \beta_0(\cdot)$, this ratio is $1/\lambda_0(t)$.). We then have the following:

$$\begin{aligned} s^{(0)}(t) &= E[Y(t)\lambda^h(t|Z)] = \int q_{Z,X,\Delta}(z,t,1)dz := q_{X,\Delta}(t,1), \\ s^{(1)}(t) &= E[ZY(t)\lambda^h(t|Z)] = \int zq_{Z,X,\Delta}(z,t,1)dz, \\ s^{(0)}(t,\beta) &= E[Y(t)\exp(Z\beta)] = \int r(z,t,\beta)q_{Z,X,\Delta}(z,t,1)dz, \quad \text{and} \\ s^{(1)}(t,\beta) &= E[ZY(t)\exp(Z\beta)] = \int zr(z,t,\beta)q_{Z,X,\Delta}(z,t,1)dz. \end{aligned}$$

The ratio $s^{(1)}(t)/s^{(0)}(t) =$

$$\int z \left\{ \frac{q_{Z,X,\Delta}(z,t,1)}{q_{X,\Delta}(t,1)} \right\} dz = E[Z|X = t, \Delta = 1].$$

At the true value $\beta_0(t)$, $s^{(1)}(t,\beta_0(t))/s^{(0)}(t,\beta_0(t)) = E[Z|X = t, \Delta = 1]$ and hence $u_w(\beta_0(t))$ is trivially zero.

Consider the following misspecified proportional hazards model scenario where $\lambda^h(t|z) = \lambda_0(t)\exp(\beta_0(t)z)$ is the true model whereas the fitted model is $\lambda(t;0)\exp(\beta z)$ with $\lambda(t;0)$ denoting the baseline hazard corresponding to the misspecified Cox model. Assume that censoring is independent of covariates so that $\bar{G}(t|z) = \bar{G}(t)$ and let $\bar{F}_1(t) = \bar{F}(t|Z=1)$ and $\bar{F}_0(t) = \bar{F}(t|Z=0)$. In addition, assume $\Pr(Z=1) = 1/2$. Then

$$s^{(0)}(t) = (1/2)\lambda_0(t)\bar{G}(t)[\bar{F}_1(t)\exp(\beta_0(t)) + \bar{F}_0(t)],$$

$$s^{(1)}(t)/s^{(0)}(t) = \bar{F}_1(t)\exp(\beta_0(t))/[\bar{F}_1(t)\exp(\beta_0(t)) + \bar{F}_0(t)], \quad \text{and}$$

$$s^{(1)}(t,\beta)/s^{(0)}(t,\beta) = \bar{F}_1(t)\exp(\beta)/[\bar{F}_1(t)\exp(\beta) + \bar{F}_0(t)].$$

Now, suppose that there is no treatment benefit in the first 6 months (say) and a strong proportional hazards benefit thereafter (hazard ratio=0.5, say) so that $\beta_0(t) = \log(0.5)I(t>6)$. Let $\theta^6 = 0.5$ denote the hazard ratio after 6 months. Note that for $t \leqslant 6$ $\bar{F}_1(t)\exp(\beta_0(t)) = \bar{F}_0(t)$, $s^{(0)}(t) = \lambda_0(t)\bar{G}(t)\bar{F}_0(t)$, and the solution β_w^* to (3.6) satisfies

$$\begin{aligned} 0 = &\int_0^6 w(t)\left\{\frac{1}{2} - \left\{\frac{\exp(\beta_w^*)}{\exp(\beta_w^*)+1}\right\}\right\}\lambda_0(t)\bar{G}(t)\bar{F}_0(t)dt \\ &+ \int_6^\infty w(t)\left\{\exp(\beta_w^*)a(t,\beta_w^*) - \theta^6\right\}(1/2)\lambda_0(t)\bar{G}(t)\bar{F}_1(t)dt, \end{aligned}$$

where $a(t,\beta_w^*) = [\bar{F}_1(t)\theta^6 + \bar{F}_0(t)]/[\bar{F}_1(t)\exp(\beta_w^*) + \bar{F}_0(t)]$ for $t > 6$. Evidently, for positive weights on $t \leqslant 6$ the first term "weights heavier towards the null", whereas for positive weights on $t > 6$ the second term "weights heavier towards θ^6".

To see how various weights perform we solve the above equation for weighting schemes FH(0,0), FH(0,1) FH(1,0), and FH(1,1). Here survival times are exponentially distributed with $\lambda_0(t) = 0.077$ for control ($Z = 0$) and $\lambda_0(t)\exp(\beta_0(t))$, with $\beta_0(t) = \log(0.5)I(t>6)$ (as above, for $Z = 1$) and exponential censoring times with hazard $\lambda_C = 0.004$. Figure 2 displays the Kaplan-Meier population curves as well as the weighting schemes.

Figure 3 displays $u_w(\beta)$ for standard Cox, the optimal weighting scheme (which is $w(t) = I(t > 6)$), FH(0,1), FH(1,0), and FH(1,1) weighting schemes. The solutions are provided in Table 1 which also provides the averages for these weighted estimators in a small sample simulation ($n = 100$ per group) to check the accuracy of the asymptotic approximation in small samples.

Here we see that the weighted estimators based on weighting schemes that focus on later differences can provide relatively accurate summaries for corresponding delayed effects.

4. Simulations

We consider combinations with the standard Cox model among FH(0,1), FH(1,0), FH(1,1), and RMST. The combinations evaluated are: Zmax1 combines Cox with FH(0,1); Zmax2 combines Cox, FH(0,1), and RMST; Zmax3 combines Cox and FH(1,0); Zmax4 combines Cox, FH(0,1), FH(1,0), FH(1,1), and RMST. Among these, Zmax1 will be most efficient for late-separation and Zmax3 most efficient for early-separation. Zmax2 attempts to cover both early and late separation by including RMST as this could be efficient in early-separation scenarios (i.e., issues such as truncation would presumably be negligible). The Zmax4 is anticipated to be most comprehensive but less efficient.

For tests that include the RMST, we define the $\hat{\beta}_{max}$ estimator to be the weighted log-rank estimator corresponding to the largest weighted log-rank z-statistic (e.g., Cox or FH(0,1) for Zmax2).

Data were generated using the nphsim R package (Wang *and others*, 2018) where we incorporate a 12 month enrollment (with ramp-up) period with administrative censoring at $t = 30$ representing the end-of-study and exponential censoring times with hazard $\lambda_C = 0.004$ representing lost-to-follow-up times. Non-proportional hazards scenarios of late-separation and early-separation are considered. The late-separation scenario is $hr(t) = 0.5I(t > 6)$ and population curves are displayed in Figure 2. Whereas for the early-separation scenario $hr(t) = 0.45$ for

$t \leqslant 10$; and $hr(t) = 1.5$ for $t > 10$ (Figure 4).

We conduct $5,000$ simulations each based on a sample size of $n = 300$ (150 per arm) wherein the synthetic-martingale resampling approach uses $1,000$ draws to calculate p-values for the combination tests Zmax1-Zmax4 and for calculating confidence intervals for the corresponding $\hat{\beta}_{max}$ Cox model estimators and Wald tests (via equation (3.5)). For the Zmax2 and Zmax4 tests the simultaneous confidence intervals include the RMST statistic within (3.5), however the Wald tests are with respect to the $\hat{\beta}_{max}$ estimator as our focus is on estimation properties for hazard ratios. For the individual Cox, FH(0,1), FH(1,0), FH(1,1), and RMST statistics we use standard large-sample approximations. Analyses were conducted when 70% of the event times were observed and thus the observed censoring proportion is 30% and the survival time observations are truncated at $t = 30$ (via end-of-study administrative censoring).

Simulation results are summarized in Table 2 for the null model, proportional hazards (PH), late-separation (LS), and early-separation (ES) scenarios.

For the null model, note that a true 2.5% size test of superiority will have a standard error of $\sqrt{0.025 * 0.975/5000} = 0.002$. We consider type-1 error rates within 2 standard errors which equals 0.029 to be 2.5%-level tests (within simulation error).

The Null model results in the upper block of Table 2 indicate that all tests adequately maintain the type-1 error rate for both the score and Wald tests including the RMST Wald type test. The confidence intervals have approximate (at least) 95% coverage for all the estimators including Zmax1-Zmax4 which are based on the simultaneous confidence interval estimation approach as previously described. The Zmax4 test has the potential for the larger bias (relative to Zmax1, Zmax2, and Zmax3 considered here) due to the selection algorithm; However, the pointwise bias is $\approx 4\%$ which appears reasonable in view of the type-1 error for the corresponding Wald test of $\approx 2.1\%$. Table 3 provides the proportion of times each of the individiual Cox, FH(0,1), FH(1,0), FH(1,1) and RMST tests are selected for the combination tests for which they are included. For

Zmax1, the test alternates between the Cox and FH(0,1) tests approximately equally, whereas for the Zmax4 test the FH(1,0) and FH(0,1) are most often selected with probabilities of approximately 39% and 35%, respectively.

For the proportional hazards (PH) model scenario (second block of Table 2) the simulation design was setup so that the log-rank test has approximately 80% power for detecting a hazard ratio of 0.68. The powers for the Zmax1-Zmax4 score tests are 74%, 73%, 76%, and 72% (resp.); Whereas the corresponding Wald test powers are 77%, 75%, 77% and 75%. In addition, the estimation bias appears limited with an average of 0.66 for the Zmax4 hazard ratio estimator compared to 0.69 for Cox; And CI coverage of (at least) $\approx$ 95% is obtained for each of the estimators. The Zmax1 and Zmax3 tests select the Cox estimator more often (76% and 68%, respectively), whereas the RMST test is more often selected for the Zmax2 and Zmax4 tests (51% and 29%).

For the late-separation (LS) scenario recall that the true hr is null prior to 6 months and 0.5 thereafter. The log-rank and Zmax3 tests have minimal power of 65% and 57% (resp.), whereas the Zmax1, Zmax2, and Zmax4 tests have powers of $\approx$ 80%. The Zmax1, Zmax2, and Zmax4 tests are primarily based on the FH(0,1) test and corresponding weighted Cox estimator with selection probabilities of 93%, 92%, and 82% (resp.). These estimators thereby focus on "potentially late effects" (via FH(0,1) weighting) and are relatively more accurate for the true effect of hr=0.5 with an average (under-estimation) bias of 0.11 (relative to the bias of 0.23 for the Cox estimator). In addition, we note that the Zmax1, Zmax2, and Zmax4 estimators have confidence interval coverage probabilities of $>$ 80% for hr=0.5, whereas the Cox and Zmax3 estimators have coverage probabilities of only 26% and 32% (resp.).

Similar results hold for the early-separation (ES) scenario as for the late-seperation scenario in terms of the influence of FH(1,0) versus FH(0,1). Here the log-rank, Zmax1, and Zmax2 tests have minimal power of 57%, 48%, and 66% (resp.), whereas the Zmax3 and Zmax4 tests have powers of 84% and 78% (resp.). The Zmax1 tests always selects the Cox model but has lower power than log-rank due to the adjustment for including the FH(0,1) test. The Zmax2 test essentially corresponds to the RMST test (selected 99.6% of the time) whereas the Zmax3 and Zmax4 test are primarily the FH(1,0) test (selected $>$ 96% of the time). We note that the Zmax2 Wald test has poor peformance relative to the Zmax2 score test because the hazard ratio estimator is based on the Cox model in this case while the increased power of the Zmax2 score test is via the RMST test.

In these simulations the Zmax2 and Zmax4 tests appear to have consistent power advantages over the log-rank test under the late-separation and early-separation NPH scenarios considered here. The Zmax4 test has the best consistent performance in terms of estimation properties (bias and CI coverage) across the scenarios considered. In addition we note that the score and Wald Zmax4 tests are in general agreement in terms of power indicating that the Zmax4 score and (adjusted) Wald confidence intervals are generally consistent with each other. Lastly, since the Zmax4 estimator will, by construction, lead to hazard ratio estimates that are "at-least as strong" as the standard Cox estimate, we note that the average difference between the Zmax4 and Cox hazard ratio estimates in the proportional hazards scenario is -0.03 (median = -0.026, 25% and 75% quantiles = -0.047, and 0); Whereas under the null, the corresponding average is -0.051 (median = -0.042, 25% and 75% quantiles = -0.071, and -0.022).

5. Data Analysis Example: Hypothetical Interim Analysis of OAK Study

For illustration we return to the hypothetical interim analysis of the OAK study where Figure 1 displays the Kaplan-Meier estimates for the treatment groups based on the July 7, 2016 cuttoff denoted "final" (Rittmeyer *and others*, 2017) and interim data based on the retrospective cutoff of June 7, 2015 denoted "interim" which is 13 months earlier. Recall that we suppose that for the total $\alpha = 1.5\%$ this was split at 0.5% for the interim and 1% for the final. Table 4 displays the

FH(ρ, γ)-weighted Cox models corresponding to the Zmax4 test where as noted the standard log-rank test would not have technically met the 0.005 interim criteria. In contrast, the Zmax4 test (adjusted) p-value is 0.0045 with companion Cox model estimated hazard ratio of $\hat{\beta}_{max} = 0.688$ (99% CI= $0.4757, 0.9946$). We note that we report the 99% (simultaneous) CI as the interim 1-sided p-value criteria is 0.5%. In addition, the RMST estimate is 1.09 (99% simultaneous CI= -0.004, 2.184) with p-value=0.002.

For the final data we note that the Zmax4 test p-value is $< 10^{-4}$ with companion weighted Cox model estimate of $\hat{\beta}_{max} = 0.699$ which for the final data corresponds to the FH$(1, 1)$ estimator and is consistent with the above hypothetical interim analysis estimate. We note that these results are also consistent with the stratified (randomization stratification factors) standard Cox hazard ratio estimate of 0.73 and p-value of 0.0003 reported in Rittmeyer *and others* (2017)

6. Discussion

In randomized clinical trial settings where treatment differences exhibit non-proportional hazards patterns, the standard log-rank and Cox model analyses can be inadequate in terms of power and estimation accuracy. For example, in a late-separation setting where there is (on average) lack of early benefit followed by substantial later benefit, the standard Cox estimator is generally neither accurate for lack of early benefit nor accurate for the later term benefit. In contrast, similar to how weighted log-rank tests can be more powerful for detecting treatment differences for which they “focus” (via the specific weighting pattern), their analogous weighted Cox model estimators can provide more accurate summaries of the effects corresponding to such differences. The weighted Cox estimators may serve as complementary summaries to the standard Cox estimator to more accurately reflect non-proportional hazards effects (e.g., the FH(0,1) estimator would reflect later term effects). In this paper we have considered FH(ρ, γ) weighted log-rank and companion Cox estimators. The synthetic-martingale resampling approach can be easily applied to combining other weighting families or combining across different weighting families (e.g., FH(ρ, γ) and Tarone-Ware families (Tarone and Ware, 1977). We note that Xu and O'Quigley (2000) consider the weight $w(t) = [1 - \hat{\bar{F}}(t-)]/\bar{Y}(t)$ and Schemper *and others* (2009) consider the weight $w(t) = [1 - \hat{\bar{F}}(t-)]/\bar{G}(t)$ where $\bar{G}(t)$ denotes the Kaplan-Meier (pooled) estimate of the censoring distribution. However, these estimators will not generally be sensitive to broad non-proportional hazards patterns (While the asymptotic limiting value may be independent of the censoring distribution, the underlying challenge of non-proportional hazards still remains.). We have also considered the inclusion of the RMST test which is a useful summary by itself. For late-separation scenarios the RMST test (due to truncation) generally loses the ability to compare differences in the later tail of the survival distributions where substantial differences could be present. We therefore considered combinations with RMST that could have advantages in early-separation scenarios where such truncation issues may not exist. In our simulations, however, combinations with the RMST test did not perform as well as combinations that included the FH(1,0) weighted log-rank test. Nonetheless it seems useful to have the RMST as an option to consider and may have preferrable performance in other settings. We note that the proposed methods based on weighted log-rank statistics can be easily extended to incorporate stratification (e.g., by randomization). An important open problem for future research is to develop interim analysis procedures. Recently Yoshida and Matsuyama (2016) have investigated interim analysis methods for weighted log-rank tests and discuss simulation approaches for evaluating their operating characteristics. The nphsim package (Wang *and others*, 2018) could be utilized in such efforts. In addition, the tools outlined in this paper for evaluating the asymptotic bias properties of candidate estimators (e.g., via equation (3.6)) and operating characteristics of test statistics may be useful guides for developing more accurate estimands in clinical trials (Phillips *and others*, 2016; Akacha *and others*, 2017).

7. Software

R code illustrated on a simulated dataset generated from the nphsim package (Wang *and others*, 2018) can be found on https://github.com/larry-leon/combination-tests-and-estimators. For questions or comments about the shared code, contact the corresponding author (larry.leon.05@post.harvard.edu).

8. Supplementary Material

8.1 *Weighted Cox model variance estimation and simultaneous confidence interval calculations*

We briefly outline some details on the weighted Cox estimator and simultaneous confidence intervals based on synthetic-martingale resampling. By Taylor series expansion

$$\sqrt{n_0+n_1}(\hat{\beta}_w-\beta_w^*)\approx\left\{i_w(\beta_w^{**})/(n_0+n_1)\right\}^{-1}(n_0+n_1)^{-1/2}U_w(\beta_w^*),$$

where β_w^{**} lies between $\hat{\beta}_w$ and β_w^*, and $i_w(\beta_w^{**})=(-1)\frac{\partial U_w(\beta)}{\partial\beta}|_{\beta=\beta_w^{**}}$ with $i_w(\beta)$ given by

$$i_w(\beta)=\int_0^\infty w(t)\left\{\frac{\bar{Y}_0(t)\exp(\beta)\bar{Y}_1(t)}{(\bar{Y}_0(t)+\exp(\beta)\bar{Y}_1(t))^2}\right\}d\bar{N}(t). \tag{8.7}$$

Define $a_j=n_j/(n_0+n_1)$ for $j=0,1$ and

$$K(t,\beta)=\sqrt{\frac{n_0+n_1}{n_0n_1}}w(t)\frac{\bar{Y}_0(t)\exp(\beta)\bar{Y}_1(t)}{\bar{Y}_0(t)+\exp(\beta)\bar{Y}_1(t)},$$

where $K(t,\beta)$ is equivalent to $K(t)$ in the weighted log-rank statistic defined in (2.1) but with $\bar{Y}_1(t)$ substituted with $\exp(\beta)\bar{Y}_1(t)$.

Recall that $\bar{M}_j(t)=\sum_{i=1}^n M_{i,j}(t)$ are martingale processes for the control ($j=0$) and experimental groups ($j=1$) with $M_{i,j}(t)=N_{i,j}(t)-\int_0^t Y_{i,j}(s)\lambda_j(s)ds$ where λ_j denotes the hazard function for group $j=0,1$. We can then write $(n_0+n_1)^{-1/2}U_w(\beta)=$

$$\sqrt{a_0a_1}\int_0^\infty \frac{K(t,\beta)}{\exp(\beta)\bar{Y}_1(t)}d\bar{M}_1(t) - \sqrt{a_0a_1}\int_0^\infty \frac{K(t,\beta)}{\bar{Y}_0(t)}d\bar{M}_0(t)$$
$$+\sqrt{a_0a_1}\int_0^\infty K(t,\beta)\frac{(\lambda_1(t)-\exp(\beta)\lambda_0(t))}{\exp(\beta)}dt.$$

We assume uniform convergence in probability for $K(\cdot,\cdot)$ and

$$\sqrt{a_0a_1}\int_0^\infty K(t,\beta)\frac{(\lambda_1(t)-\exp(\beta)\lambda_0(t))}{\exp(\beta)}dt \to \zeta(\beta) \quad \text{(say)}.$$

In addition, let π_j denote the limits of a_j for $j=0,1$. Then, $(n_0+n_1)^{-1/2}U_w(\beta_w^*) \approx$

$$\sqrt{\pi_0\pi_1}\int_0^\infty \frac{K(t,\beta_w^*)}{\exp(\beta_w^*)\bar{Y}_1(t)}d\bar{M}_1(t) - \sqrt{\pi_0\pi_1}\int_0^\infty \frac{K(t,\beta_w^*)}{\bar{Y}_0(t)}d\bar{M}_0(t) + \zeta(\beta_w^*). \tag{8.8}$$

Now, the first two terms are of the form $\int_0^\infty H_1 d\bar{M}_1(t) - \int_0^\infty H_0 d\bar{M}_0(t)$ where $H_1 = \sqrt{\pi_0\pi_1}\frac{K(t,\beta_w^*)}{\exp(\beta_w^*)\bar{Y}_1(t)}$ and $H_0 = \sqrt{\pi_0\pi_1}\frac{K(t,\beta_w^*)}{\bar{Y}_0(t)}$. Let h_1 denote the limit of $H_1^2\exp(\beta_w^*)\bar{Y}_1$ and h_0 denote the limit of $H_0^2\bar{Y}_0$. Then following arguments as in Fleming and Harrington (1991) (See page 268) the variance of $(n_0+n_1)^{-1/2}U_w(\beta_w^*)$ can be estimated by (upon plugging in empirical counterparts for limits) $\hat{\sigma}^2(U_w(\beta_w^*)) =$

$$(\pi_0\pi_1)\int_0^\infty \left\{\frac{K(t,\beta_w^*)^2}{\bar{Y}_0(t)} + \frac{K(t,\beta_w^*)^2}{\exp(\beta_w^*)\bar{Y}_1(t)}\right\}\left\{1 - \frac{\Delta\bar{N}(t)-1}{\bar{Y}_0(t)+\exp(\beta_w^*)\bar{Y}_1(t)-1}\right\}$$
$$\times \frac{d\bar{N}(t)}{\bar{Y}_0(t)+\exp(\beta_w^*)\bar{Y}_1(t)}. \tag{8.9}$$

Note that for $\beta_w^* = 0$ the score $U_w(0)$ and $\hat{\sigma}^2(U_w(0))$ are equivalent to the score statistic (2.1) and its variance estimate (2.2). The variance of $\hat{\beta}_w$ is estimated by plugging-in $\hat{\beta}_w$ in equations (8.7) and (8.9).

Now, for calculating the constant c_{max} in the simultaneous confidence interval described in (3.5), we use

$$\sqrt{n_0+n_1}(\hat{\beta}_w - \beta_w^*) \approx \left\{i_w(\hat{\beta}_w)/(n_0+n_1)\right\}^{-1}(n_0+n_1)^{-1/2}U_w^\dagger(\hat{\beta}_w),$$

where in analogy to (2.3) $N_{i,j}G_{i,j}$ terms are plugged-in for the $M_{i,j}$'s in expression (8.8) and the same $G_{i,j}$'s would be used for each weighted estimator within (3.5).

Acknowledgments

The authors are grateful to the Atezolizumab Lung Team as well as Zhengrong Li, Kaspar Rufibach, and Marcel Wolbers for their helpful comments and suggestions.

Conflict of Interest: None declared.

References

Akacha, Mouna, Bretz, Frank, Ohlssen, David, Rosenkranz, Gerd and Schmidli, Heinz. (2017). Estimands and their role in clinical trials. *Statistics in Biopharmaceutical Research* **9**(3), 268–271.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**(4), 1100–1120.

Chi, Yunchan and Tsai, Min-Hsiao. (2001). Some versatile tests based on the simultaneous use of weighted log-rank and weighted kaplan-meier statistics. *Communications in Statistics - Simulation and Computation* **30**(4), 743–759.

Dobler, D., Beyersmann, J. and Pauly, M. (2017). Non-strange weird resampling for complex survival data. *Biometrika* **104**(3), 699–711.

Eng, Kevin Hasegawa and Kosorok, Michael R. (2005). A sample size formula for the supremum log-rank statistic. *Biometrics* **61**(1), 86–91.

Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*, Wiley Series in Probability and Statistics. Wiley.

Fleming, Thomas R., Harrington, David P. and O'sullivan, Margaret. (1987). Supremum versions of the log-rank and generalized wilcoxon statistics. *Journal of the American Statistical Association* **82**(397), 312–320.

Gill, R.D. (1980). Censoring and stochastic integrals. *Statistica Neerlandica* **34**(2), 124–124.

Gill, Richard. (1983, 03). Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist.* **11**(1), 49–58.

Goldwasser, M. A., Tian, L and Wei, L. J. (2004). Statistical inference for infinite-dimensional parameters via asymptotically pivotal estimating functions. *Biometrika* **91**, 81–94.

Kosorok, Michael R. and Lin, Chin-Yu. (1999). The versatility of function-indexed weighted log-rank statistics. *Journal of the American Statistical Association* **94**(445), 320–332.

Lee, Seung-Hwan. (2007). On the versatility of the combination of the weighted log-rank statistics. *Computational Statistics Data Analysis* **51**(12), 6557 – 6564.

Lin, D. Y. (1991). Goodness-of-fit analysis for the cox regression model based on a class of parameter estimators. *Journal of the American Statistical Association* **86**(415), 725–728.

Lin, D. Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine* **16**(8), 901–910.

Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

Pepe, Margaret Sullivan and Fleming, Thomas R. (1989). Weighted kaplan-meier statistics: A class of distance tests for censored survival data. *Biometrics* **45**(2), 497–507.

Phillips, Alan, Abellan-Andres, Juan, Soren, Anderson, Bretz, Frank, Fletcher, Chrissie, France, Lesley, Garrett, Andrew, Harris, Raymond, Kjaer, Magnus, Keene, Oliver, Morgan, David, O'Kelly, Michael *and others*. (2016). Estimands: discussion points from the psi estimands and sensitivity expert group. *Pharmaceutical Statistics* **16**(1), 6–11.

Rittmeyer, Achim, Barlesi, Fabrice, Waterkamp, Daniel, Park, Keunchil, Ciardiello, Fortunato, von Pawel, Joachim, Gadgeel, Shirish M, Hida, Toyoaki, Kowalski, Dariusz M, Dols, Manuel Cobo, Cortinovis, Diego L, Leach, Joseph, Polikoff, Jonathan, Barrios, Carlos, Kabbinavar, Fairooz, Frontera, Osvaldo Arn, De Marinis, Filippo, Turna, Hande, Lee, Jong-Seok, Ballinger, Marcus, Kowanetz, Marcin, He, Pei, Chen, Daniel, Sandler, Alan *and others*. (2017). Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (oak): a phase 3, open-label, multicentre randomised controlled trial. *The Lancet* **389**(10066), 255–265.

Royston, Patrick and Parmar, Mahesh KB. (2016). Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC medical research methodology* **16**(1), 16.

Sasieni, Peter. (1993). Maximum weighted partial likelihood estimators for the cox model. *Journal of the American Statistical Association* **88**(421), 144–152.

Schemper, Michael, Wakounig, Samo and Heinze, Georg. (2009). The estimation of average hazard ratios by weighted cox regression. *Statistics in Medicine* **28**(19), 2473–2489.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73**(2), 363–369.

Tarone, Robert E. and Ware, James. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**(1), 156–160.

Tian, L., Liu, J., Zhao, M. and Wei, L. J. (2004). Statistical inferences based on non-smooth estimating functions. *Biometrika* **91**, 943–954.

Wang, Yang, Wu, Haiyan and Anderson, Keaven. (2018). NPHSIM: Simulation and power calculations for time-to-event clinical trials. *https://github.com/keaven/nphsim/*.

Xu, Ronghui and O'Quigley, John. (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics* **1**(4), 423–439.

Yoshida, Mizuki and Matsuyama, Yutaka. (2016). Interim analysis based on the weighted log-rank test for delayed treatment effects under staggered patient entry. *Journal of Biopharmaceutical Statistics* **26**(5), 842–858. PMID: 26391147.

Zhao, Lihui, Claggett, Brian, Tian, Lu, Uno, Hajime, Pfeffer, Marc A., Solomon, Scott D., Trippa, Lorenzo and Wei, L. J. (2016). On the restricted mean survival time curve in survival analysis. *Biometrics* **72**(1), 215–221.

Table 1. Limiting values of weighted hazard ratio estimators

	Cox	FH(0,1)	FH(1,0)	FH(1,1)	Opt
Asymptotic	0.669	0.563	0.770	0.630	0.50
Simulations	0.698	0.587	0.789	0.645	0.51

[]

Table 2. Properties of tests and hazard ratio estimators under null, proportional hazards (PH), late-separation (LS), and early-separation (ES) scenarios. Monte carlo summaries are defined as: Est denotes the empirical average of the estimated hazard ratios; SE denotes the empirical standard errors for the hazard ratio estimates; ESE denotes the empirical average of the estimated standard errors; CP(null) denotes the coverage probability of a null effect ($hr = 1$ for hr estimates and 0 for RMST), whereas CP(0.68), CP(0.5), and CP(0.45) denote the coverage probabilities under the PH, LS, and ES scenarios (e.g., CP(0.68) is the coverage probability for the true PH effect of hr=0.68).

	Cox	Zmax1	Zmax2	Zmax3	Zmax4	FH(0,1)	FH(1,0)	FH(1,1)	RMST
Null									
Score	0.0262	0.0204	0.0178	0.0206	0.0182	0.0272	0.0208	0.0284	NA
Wald	0.0254	0.0248	0.0218	0.0228	0.0208	0.0270	0.0208	0.0278	0.0252
Est	1.010	0.980	0.980	0.992	0.958	1.014	1.010	1.011	0.000
SE	0.142	0.147	0.147	0.142	0.140	0.166	0.149	0.152	0.921
ESE	0.140	0.147	0.147	0.141	0.145	0.163	0.147	0.150	0.910
CP(null)	0.949	0.965	0.969	0.965	0.975	0.947	0.956	0.948	0.949
Proportional hazards (PH)									
Score	0.795	0.742	0.733	0.755	0.724	0.676	0.750	0.749	NA
Wald	0.793	0.766	0.746	0.773	0.754	0.671	0.750	0.746	0.787
Est	0.686	0.670	0.670	0.675	0.656	0.688	0.686	0.687	2.946
SE	0.0971	0.0981	0.0981	0.0961	0.0948	0.1135	0.1016	0.1039	1.0995
ESE	0.0955	0.0968	0.0968	0.0957	0.0966	0.1113	0.1006	0.1023	1.0538
CP(0.68)	0.948	0.966	0.969	0.963	0.974	0.945	0.954	0.946	NA
Late-separation (LS)									
Score	0.647	0.811	0.798	0.567	0.778	0.872	0.357	0.789	NA
Wald	0.645	0.827	0.815	0.584	0.796	0.870	0.356	0.788	0.571
Est	0.728	0.614	0.614	0.728	0.614	0.612	0.802	0.665	2.417
SE	0.105	0.102	0.102	0.105	0.102	0.102	0.121	0.104	1.205
ESE	0.1015	0.0994	0.0994	0.1015	0.0988	0.1003	0.1172	0.0999	1.1118
CP(null)	0.355	0.173	0.185	0.416	0.204	0.130	0.643	0.212	0.429
CP(0.50)	0.264	0.817	0.829	0.321	0.844	0.777	0.126	0.549	NA
Early-separation (ES)									
Score	0.5760	0.4762	0.6574	0.8350	0.7778	0.0354	0.8796	0.2288	NA
Wald	0.5736	0.5056	0.4678	0.8496	0.7998	0.0346	0.8782	0.2264	0.7720
Est	0.749	0.749	0.749	0.636	0.636	1.009	0.636	0.853	2.427
SE	0.1079	0.1079	0.1079	0.0952	0.0952	0.1734	0.0952	0.1354	0.8946
ESE	0.1040	0.1040	0.1040	0.0936	0.0936	0.1650	0.0936	0.1269	0.8911
CP(null)	0.426	0.494	0.532	0.150	0.200	0.942	0.122	0.772	0.228
CP(0.45)	0.0596	0.0784	0.0954	0.4238	0.5086	0.0046	0.3798	0.0188	NA

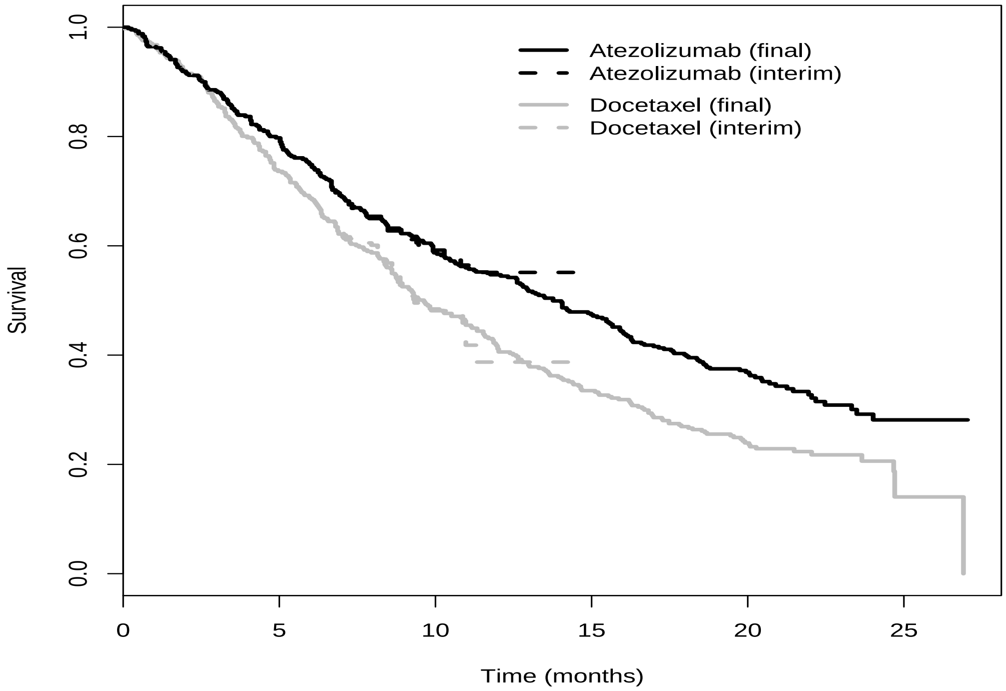

Fig. 1. OAK Kaplan-Meier overall survival estimates at the final (solid curves) and interim analyses (dashed curves).

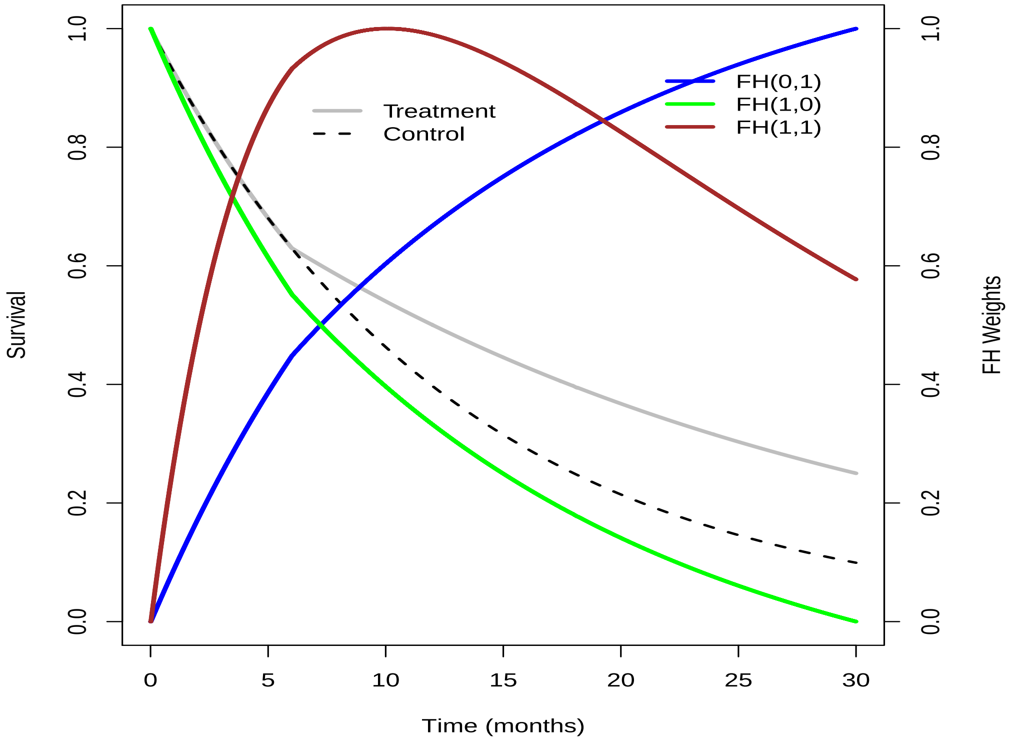

Fig. 2. Underlying Kaplan-Meier curves under late-separation scenario along with FH(0,1), FH(1,0) and FH(1,1) weight profiles.

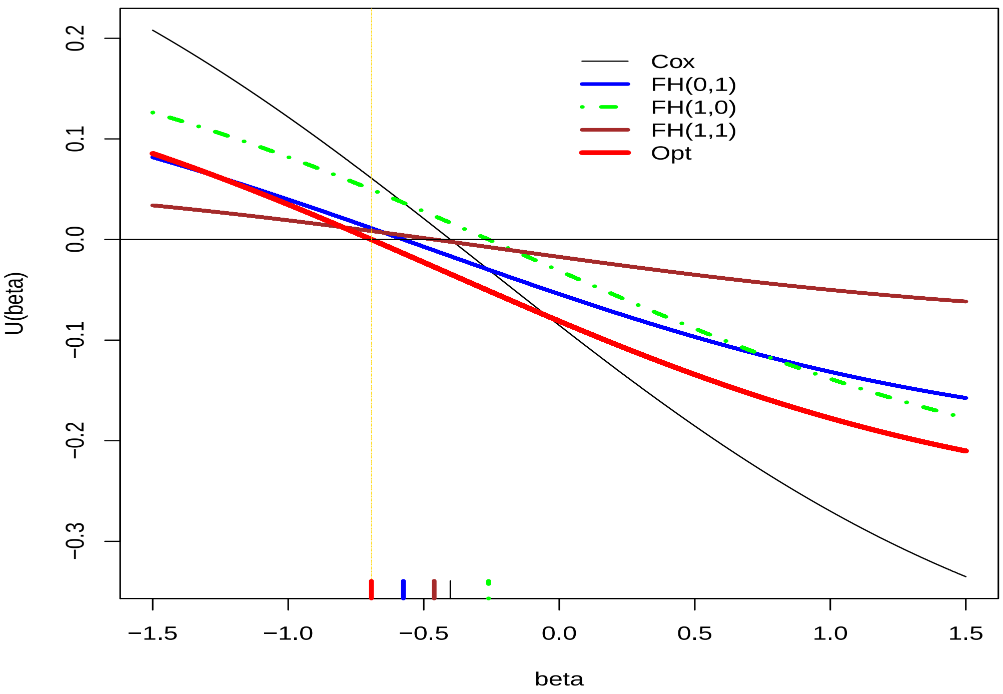

Fig. 3. Solutions to $u_w(\beta)$ for weight functions: FH(0,0) the standard Cox model (denoted Cox), FH(0,1), FH(1,0), FH(1,1), and Opt which denotes the "optimal" weight function $w(t) = I(t > 6)$.

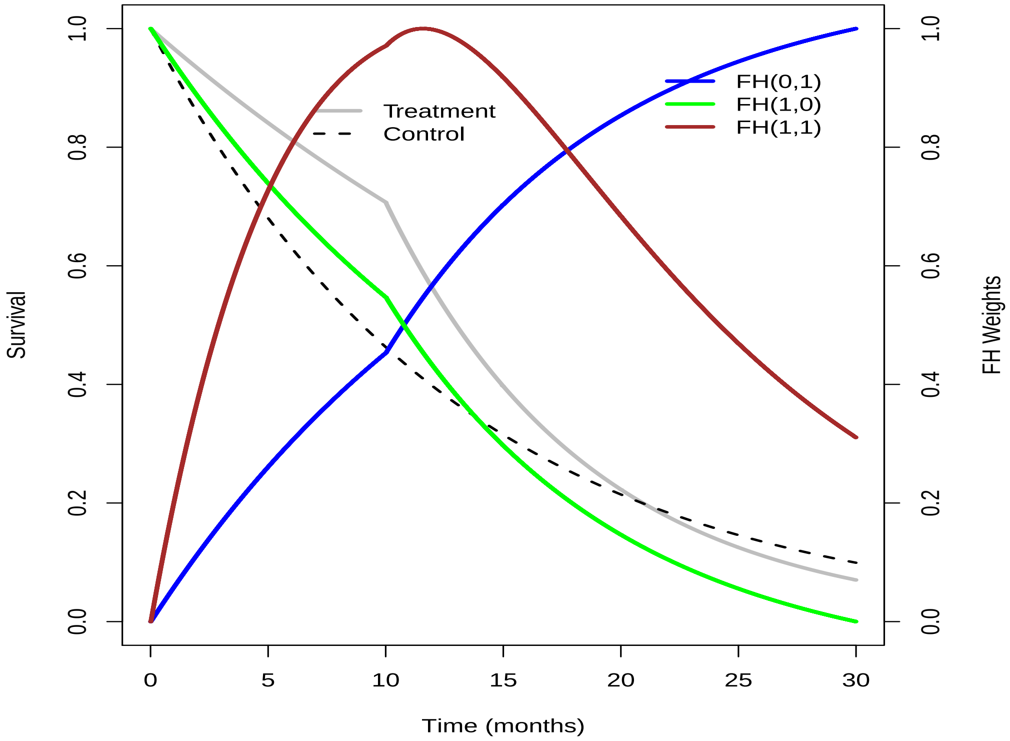

Fig. 4. Underlying Kaplan-Meier curves early-separation scenario along with FH(0,1), FH(1,0) and FH(1,1) weight profiles.

Table 3. The proportion of times each of the individual tests (Cox, FH(0,1), FH(1,0), FH(1,1), and RMST) correspond to the maximizer of the combination tests: Zmax1 combines Cox with FH(0,1); Zmax2 combines Cox, FH(0,1), and RMST; Zmax3 combines Cox and FH(1,0); Zmax4 combines Cox, FH(0,1), FH(1,0), FH(1,1), and RMST.

	Cox	FH(0,1)	FH(1,0)	FH(1,1)	RMST
Null					
Zmax1	0.5032	0.4968	-	-	-
Zmax2	0.1272	0.476	-	-	0.3968
Zmax3	0.4924	-	0.5076	-	-
Zmax4	0.0192	0.3546	0.3894	0.1718	0.065
Proportional hazards (PH)					
Zmax1	0.763	0.2374	-	-	-
Zmax2	0.252	0.2346	-	-	0.5138
Zmax3	0.680	-	0.3204	-	-
Zmax4	0.160	0.1616	0.2142	0.1794	0.2852
Late-separation (LS)					
Zmax1	0.0734	0.9266	-	-	-
Zmax2	0.0432	0.924	-	-	0.0328
Zmax3	0.9950	-	0.005	-	-
Zmax4	0.0304	0.8218	0.0028	0.125	0.02
Early-separation (ES)					
Zmax1	1.0000	0	-	-	-
Zmax2	0.0038	0	-	-	0.9962
Zmax3	0.0020	-	0.998	-	-
Zmax4	0.0002	0	0.9662	0.001	0.0326

Table 4. OAK 'hypothetical interim' analysis

					Pointwise 99% CI		Simultaneous 99% CI	
	HR	SE	Z	pvalue	Lower	Upper	Lower	Upper
Cox	0.754	0.083	2.575	0.00501	0.5687	1.0009	0.5520	1.0312
FH(0,1)	0.688	0.089	2.912	0.00179	0.4927	0.9602	0.4757	0.9946
FH(1,0)	0.773	0.086	2.316	0.01027	0.5808	1.0298	0.5635	1.0613
FH(1,1)	0.706	0.086	2.859	0.00212	0.5150	0.9680	0.4982	1.0007